Volume 5, Number 11
November 1991

Is Resource Discovery Hacking?

The Great Measurement Debacle

by Carl Malamud

Introduction

Mike Schwartz is an Assistant Professor at the University of Colorado. He specializes in an emerging area of network applications known as *resource discovery*. Instead of some purely academic pursuit such as reinventing the remote procedure call, Schwartz tries to conduct research on the existing Internet and how to use it more effectively.

Resource Discovery

There are several examples of resource discovery. Schwartz has developed software called *netfind*, which attempts to locate electronic mail addresses. Rather than using a single source of information, *netfind* uses *finger*, SMTP, DNS, and a variety of other sources to try to locate users. The Corporation for National Research Initiatives has developed a system similar to *netfind*, known as the *Knowbot Information Service* (KIS) and is working on intelligent search techniques for information archives.

The *Archie* project at McGill University compiles a list of files available via Anonymous FTP. *Archie* allows the user to discover where in the Internet a file exists. Other projects include *Prospero* developed by Clifford Neuman of ISI and *WAIS*, pioneered by Brewster Kahle of Thinking Machines.

Study

A recent series of incidents underscores many of the issues in this emerging area. Schwartz was working on a research project to measure the nature of connectivity to the Internet. The study is quite simple. Using a statistical sample of hosts on the Internet, Schwartz tries to see what services they offer.

This study measures, in a statistical fashion, the nature of connection to the Internet. As a longitudinal study, Schwartz hopes to see if the nature of that connection changes over time. For example, people might start disconnecting themselves because of security concerns, as Dr. David Clark of MIT has predicted.

Services

The study is interesting in several respects. First, it is not based on simulation or other theoretical models: it uses the real Internet. His study starts with a database of some four hundred thousand hosts in over 12,000 different domains. Then, Schwartz picks a sample of that population, doing on average 3.9 name lookups per domain and attempting 5.7 connections per domain. The study attempts to open the following ports:

Port Number	Service
13	daytime
15	netstat
21	FTP
23	Telnet
25	SMTP
53	DNS
79	Finger
111	portmap
513	login
540	uucp daemon
543	klogin
544	krcmd kshell

The services were chosen carefully to try to yield information about the type of connectivity a particular host offers.

For example, ports 543 and 544 are *Kerberos*-based remote access procedures and indicate the presence of a secure gateway. The services chosen were those that one can expect to be run on every machine in a domain, thus making a statistical analysis valid. For example, one could expect SMTP on every port, but specialized protocols such as Z39.50 would probably run only on selected servers.

Only TCP ports were picked since with UDP there is no application-independent way to see if a UDP-based service is running. With TCP, one can simply see if a connection attempt succeeds to discover if the service is running. Note that successfully opening a TCP connection does not mean that the service itself will be used. Most services have some form of access control, such as a password. The Schwartz study immediately disconnects the TCP connection and never establishes an association at the application level.

The program is quite careful not to overload either individual machines or the Internet. No more than 3 connection attempts are made on no more than 3 machines in a particular domain. The software uses 20 concurrent threads, ensuring that no more than 20 probes are active at any one time on the network. The program takes roughly a day to run and contributes, on the day it is run, an increase in Internet traffic of roughly 0.5%

Sound harmless enough? Small-scale versions of the study were run in August, 1990 and February, 1991 on the *finger* and SMTP ports as a way of determining the potential scope of the *netfind* tool. A few astute system managers noted the probes and contacted the *Computer Emergency Response Team* (CERT) and University of Colorado system managers, both of whom had been advised before each of the studies.

In August, 1991, Schwartz expanded the study to look at more services. He began running his study and, again, a few managers contacted the CERT and system managers at the University of Colorado to find out what was going on. In all cases, the resolution was pretty simple. Schwartz explained his study and people said "fine." However, because the study was bigger and there were more notifications to the CERT, Schwartz volunteered to advise people about his activities.

Notifications

Two methods were used to notify system managers. First, mail was sent to the postmasters in the domains affected. Mail started to get sent, but after 4,000 messages went out it was evident that less than half of the domains had a valid postmaster account. The remaining half of the messages were undeliverable, served by an automatic response, or otherwise did not succeed in reaching a postmaster.

Schwartz also posted messages on USENET to advise people of what he was doing. Schwartz posted a message on `alt.security` that began:

> "I am in the midst of conducting a series of experiments designed to measure changes in service level reachability in the global Internet, to help characterize the extent to which institutions are distancing themselves from the Internet..."

The note went on and explained the methodology, the fact that the CERT was aware of the activity, and offered to make more information available. Messages started coming back from a few paranoid system administrators threatening to put Schwartz in jail, to call the FBI, and to otherwise cause severe problems if he attempted to open a connection on any of their hosts.

continued on next page

Is Resource Discovery Hacking? *(continued)*

Strong reactions

A large computer company, for example, explained that they were a "commercial user on the Internet" and that any attempt to conduct the study would "be treated as an intrusion and will be considered a prosecutable offense which we will pursue."

One of the most vociferous messages came from a systems manager at a startup well-known for making PC clones, who explained to Schwartz:

> "If you were to traverse our gateway(s), I'd simply call the FBI. You're way out of line, and really asking for significant legal problems. I in *no way* support what you're starting to do, it's wrong and UNETHICAL TO ENTER OTHER SITES, FOR WHATEVER 'REASON' YOU'RE TRYING TO YOURSELF JUSTIFY IT/THIS FOR." [sic]

(An interesting note, by the way, is to look up this company in the Domain Name Service. One sees that they use UUCP and UUnet to receive their mail, are not part of the Internet, and would not have been hit by this study.)

Whew! That's enough to make most researchers sit up and take notice. Lawyers still cost money, after all. With the threats of law suits coming in, Schwartz decided to temporarily table the study.

Objections

The next step was to reevaluate if this study had somehow violated the boundaries of what one is allowed to do on the Internet. Several arguments were raised in objection to the study:

- The study was a waste of machine resources
- The study was a waste of Internet bandwidth
- The study was a waste of network manager bandwidth by making them track down the intrusion and see if it was legitimate
- The study was a security violation

The first two arguments are really red herrings. The amount of machine resources was extremely small, and the load on the Internet was not substantial. The study was carefully engineered to make sure that a particular name server was not hit too many times, that a particular domain was only hit a few times, and that, in the case of failure, the study would simply give up and move on rather than keep trying.

The last two arguments, however raise fundamental issues. First, there is the question of the network manager. Several astute network managers noticed repeated tries to connect to ports on their machines and suspected an automated breakin attempt was in progress. They spent time trying to determine if they were under attack.

One of the outcomes of the Schwartz study was a realization that there needs to be some way of making system managers aware of this type of work. The `alt.security` posting didn't appear to reach a wide audience. A full-fledged CERT alert would be inappropriate for this activity. Some form of background information distribution is needed for activities of this sort.

The fundamental question is whether or not this type of work is allowed. The Internet Measurement study is one of an emerging class of applications.

People are talking about navigating the network with a joystick, searching FTP resources for relevant information in a network-wide search, and a wide variety of other applications. Schwartz may be the first to try this particular massive *ping,* but he certainly will not be the last.

It is widely accepted that a single user can use the *finger* service to look for a single user on a single machine. Is it acceptable to methodically use *finger* to try and build a directory?

What about methodically using *finger* to conduct research? Do we make a differentiation between research and commercial use? Between statistical analysis and the preservation of individually identifiable data? Do we care what type of user is conducting the probe?

Issues

There are three issues that can be raised by this type of work:

- A definition of the security perimeter on the Internet
- Privacy constraints
- Scaling

Most would agree that you are allowed to FTP to a host and attempt to login with the username "anonymous." This is a service, if it is there, that is available to the general public. Attempting to use anonymous FTP, even if the service is refused, is not a violation of the security perimeter. On the other hand, it is generally accepted that trying other usernames at random would in fact be a violation of the security perimeter.

By that definition, what Schwartz was doing did not violate the security perimeter of the Internet. He simply attempted to open a so-called *well-known port.* He did not try to login or to conduct any sort of attack on the perimeter.

Internet reaction

The opinion of many members of the Internet community was solicited on this issue. Almost unanimously, IAB, IESG, and NSF officials responded in a similar manner: "So what's the problem?" A typical response was from Dr. David D. Clark of MIT:

> "In general, the community has considered it out of bounds to attack systems, even with no malice, to see if they have a security flaw. We consider it in bounds to touch a system (e.g., *finger*) to see if it is there. So, without reading his info files, I would conclude that this experiment is within the bounds of acceptable behavior."

If a service is available to the public, such as the *finger* service, one is able to use the service. By this analysis, attempting to open TCP ports is perfectly legitimate. Trying to hack a username/password is not.

If you have a well-known port on the Internet, you should be prepared for people to attempt a connection. Using this analysis, the Schwartz study was perfectly legitimate. The activity he was conducting would have been perfectly legitimate if done manually and the fact that a program was conducting the activity in a wholesale fashion does not change the security issues.

Secure gateways

Many institutions have begun using secure Internet gateways. These gateways protect the hosts inside of the domain from exactly this type of intrusion. In other words, organizations that don't want to be subject to research like that of Mike Schwartz ought to install a secure Internet gateway. This is exactly what most major computer companies, including Sun, DEC, and IBM, have done.

continued on next page

Is Resource Discovery Hacking? *(continued)*

Privacy

Privacy falls within the same boundaries. Organizations that wish to protect the privacy of their users turn off services that violate that privacy. For example, many sites choose not to offer the *finger* service to the outside world. The Schwartz study would, in this case, attempt to open the *finger* port and record a failure, indicating that this particular site is not offering the *finger* service. If three hosts in a domain all fail to offer the service, the study then assumes that the service is not offered in the domain.

It is important to realize that we need both protected and unprotected forms of connection to the Internet. The Internet is a global community and simply prohibiting people from walking the streets of the global village is not an adequate solution. If you don't want people looking in your window, then pull the shades.

There is an aspect of privacy that does need to be considered, however. Some information, based on a single query, might not violate privacy. For example, given a real name, we can get an electronic mail address. However, one nature of a computer network is that we can make repeated narrow queries to amass large quantities of information. This is one of the fears of offering X.500 services: some people might make use X.500 to make copies of an entire directory for marketing or other purposes rather than finding narrow pieces of information.

Scaling

The Schwartz study does raise an important question when it comes to scaling: even if we allow a single Schwartz to conduct research what happens when 1,000 high school students begin emulating him? One person *pinging* a port is not a problem: millions of people doing so certainly is. Another example of this issue is the *Archie* project. One *Archie* server getting a directory listing off your system is not a drain on resources, but one million *Archies* doing so would be.

We need to support resource discovery on the Internet, but we also need to think very carefully about setting up an infrastructure to support this class of applications. This is a crucial area of research. A single centralized directory will not solve people's desire for information. Different users need different kinds of information and a resource discovery infrastructure needs to support a wide variety of different classes of modules, ranging from simple IP *pings* to Knowbots.

How to solve this type of problem is currently under study by an Internet Research Task Force research group headed by Schwartz. The research group is focusing on the technical issues involved in resource discovery. The IETF has also made a few stabs at resource discovery, but it is evident that we need to know a lot more about this class of applications before any hit the RFC stage.

Conclusion

The need to learn more is one of the most compelling reasons to allow Schwartz to move forward on this type of application. If qualified researchers are not able to expand the functionality of the Internet, we will cease to make progress. Only by seeing what happens when these types of experiments are being conducted can we begin to think about an Internet-wide infrastructure to support resource discovery.

For further reading

- Papers written by Mike Schwartz can be obtained by anonymous FTP from `latour.cs.colorado.edu`. See also the article by Schwartz in the May 1991 issue of *ConneXions* (Volume 5, No. 3).
- *Knowbots* are described in Malamud, *STACKS—The INTEROP Book,* (Prentice Hall, 1991). To access the Knowbot Information Service, send mail to `KIS@NRI.Reston.Va.US` and put "?" in the first line of the message.
- WAIS is described in this issue of *ConneXions,* see page 2–9.
- *Archie* can be accessed by Telnet at `Quiche.CS.McGill.CA`. Login as "Archie," no password is required.
- *Prospero* can be retrieved by anonymous FTP from Internet host `cs.washington.edu`.

CARL MALAMUD (`carl@malamud.com`) works with Mike Schwartz on issues of resource discovery. Malamud is the author of several books, including *STACKS—The INTEROP Book.* He is currently traveling around the world doing research for a technical travelogue.

The Internet Gopher: A Distributed Information Service

What is it?

The *Internet Gopher* is a distributed document delivery service. It allows a neophyte user to access various types of data residing on multiple hosts in a seamless fashion. This is accomplished by presenting the user a hierarchical arrangement of documents and by using a client-server communications model. The Internet Gopher Curses Client allows a user on a terminal to access the vast array of information available on various gopher servers. The Internet Gopher Server accepts simple queries, and responds by sending the client a document.

Software

Also included in the release are experimental clients and servers for real-time radio, utilities for using gopher in shell scripts (written in *perl*), and some sound utilities for NeXT machines. Other Internet Gopher Software available includes:

- Macintosh Gopher client written in *HyperCard.*
- Macintosh Gopher Server software.
- PC Gopher Client with a Borland *Turbo Vision* interface.
- Full Text Indexing servers for NeXT machines.
- NeXT Gopher client (provided by Max Tardiveau of the University of St. Thomas.)

Getting a copy

All of this software is available for anonymous FTP from `boombox.micro.umn.edu` (128.101.95.95) in the `/pub/gopher` directory.

The Internet Gopher Development Team can be reached via e-mail as `gopher@boombox.micro.umn.edu`. If you prefer paper we can be reached at:

Internet Gopher Team
Computer & Information Services
University of Minnesota
132 Shepherd Labs
100 Union Street SE
Minneapolis, MN 55455